# Testing Analytic Techniques

m.hill@cranfield.ac.uk / martin.hill@armymail.mod.uk

We usually have too little time and too much data to assess, communicate and act before the situation changes. The relevant data "SUCs": it is Scattered across different sources, Unreliable, and Cluttered by irrelevant data [1]. There is too much for one brain to cope with, so we work as teams – our cognition is distributed across individuals and groups – with all the extra time demands and cognitive loads that brings.

As information systems grow and develop they make even more data available, and as the tempo of operations increases we have even less time to analyse it. We therefore need to be correspondingly better at identifying what we need and analysing what we have to, in order to produce useful assessments for decision makers to take useful actions.

We have some tools already. The NATO/Admiralty codes [2] provide us with ways to describe uncertainty in sources. Kent gave us terminology to specify uncertainty [3] in our assessments. Pherson & Heuer [4], amongst others [5] [6], have identified and described some techniques we can use to be better at extracting the right data, sharing and assessing it as groups, and presenting it to decision makers.

But do these tools work?

Anecdotes from operations suggest these tools are not used as intended. Few analysts use Alternative Competing Hypotheses (ACH) formally, either as individuals or groups, because 'it takes too long'. NATO codes are usually collapsed to only A1 [good], C3 [iffy]and F6 [bad/unknown], losing the rich possibilities of the code, conflating the very different meanings of 'bad' with 'don't know', and confusing confidence with probability. Words of Estimative Probability tend to be collapsed to 'unlikely', 'possible' and 'highly likely'.

Why is this?

Perhaps the costs in time are too high; if formal processes such as ACH take 'too long' we should not be surprised if analysts under time pressure discard such clumsy tools. Perhaps the cognitive load becomes too high; when dealing with lots of data, *adding* metadata such as confidence markers increases complexity enough to make the situation intractable.

Perhaps our tools might not be fit for purpose. When we have time we can sit back and think about what we need and invent ways to be better, but then when we sit forward and get back to work the tools we invented don't work as expected.

This suggests that we need to test our tools; we need to understand what the trade offs are. Do the tools trade time for objectivity or accuracy? Are the tools suitable but used inappropriately? Are they simply a distraction – do they consume attention without benefit? Or, worse, do we believe our products are better when using tools when actually they are worse?

Some attempts have been made to test them. An experimental test of ACH [7] concluded that it made little difference to the intelligence assessments produced, but the test was based on a trivial scenario rather than the complex ones that ACH is intended for. 'Critical assessment tests' (CATs) are supposed to evaluate students' assessment skills, but Possin's review [8] of a variety of commercial and academic CATs found few of them even tested what they claimed to test.

It seems we rarely test our tools to see if they are fit for purpose, and when we do even our tests are not fit for the purpose of testing them.

Part of the problem, as touched on by Possin, is that if we test on real world problems we can only compare the subject's assessment with the experimenter's assessment. The underlying truth of the real world is never properly known for intelligence work, we only have reports and assessments based on them, and these cannot therefore be truly objective. If the experiment designers declare that they *know* what *should* have been the conclusion, and then judge the subjects against *their* assessment, they are merely declaring that the experimenters are better analysts.

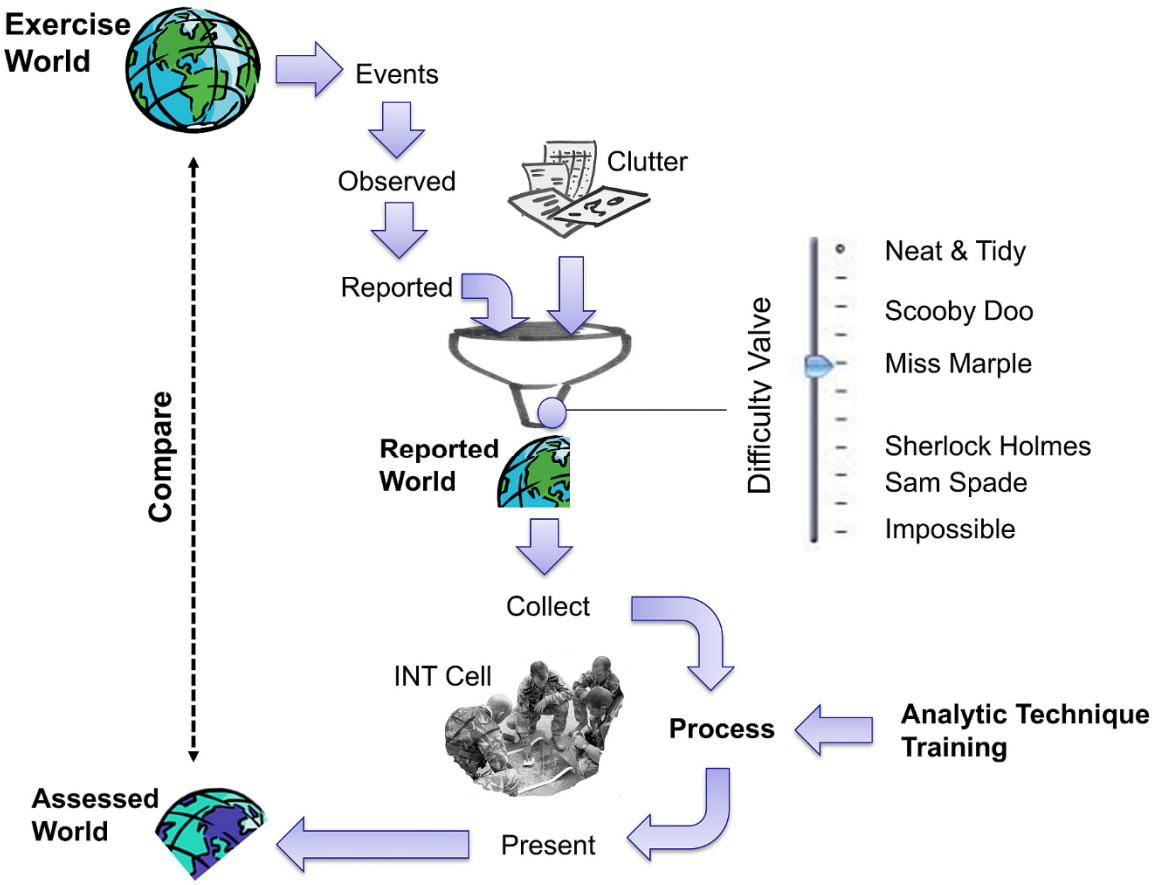
So how should we test our tools?

I suggest we create Exercise Worlds that provide a 'base truth'. They are realistically complex and involved, but simple enough for experimental assessors to comprehend. They are stories of motivated people who act: acquire, persuade, assess, plan, move, attack, defend.

These actions can be observed and reported, and the reports become the 'inserts' into the exercise. The reports are a subset of the Exercise World – this is the Reported World.

The test subjects analyse the Reported World that they have been given, assess the situation, and then describe their Assessed World.

Experimenters compare the Assessed World against the Exercise World and the differences between the Assessed World and Exercise World can be correlated with the differences in the analytic techniques used.



This is not far different from existing training practice (giving students a simple exercise world to assess), but underlaid with the full story, and overlaid with an experimental test framework.

For example, the Experimental World could consist of a DATE-based scenario of an advancing Arianan Division, described down to company level with all the associated general and specialist recce and logistics efforts. A subset of the story would be selected to be the actions observed by friendly forces, and are the inserts for battle tracking experiment subject teams. Some subject teams would be trained in ACH to evaluate options and some not. The teams would produce assessed positions, units, objectives and timings – the assessed world. These would be compared with the exercise positions, units, objectives and timings – the exercise world.

This is not a clean, solid-science test; there are many variables between experiments, some of them are human, and many are poorly described. Measuring the 'closeness' of the Assessed World with the Exercise World will require combining several dimensions and topics. We cannot run the same test on the same team more than once, and there are likely large differences in competence, experience and aptitudes between teams. Ambiguity in the Reported World can result in valid assessments (in that they match the reports) but large discrepancies with the Exercise World.

Nevertheless other research domains, including medicine and the more 'solid' social sciences, have already faced these issues and to some extent have developed experimental protocols to remove or mitigate these issues.

The key is to introduce this introspective testing of our own processes in order to establish what the trade-offs are between time, cost, training, accuracy, objectivity, rigour and so on. Once we know what the trade-offs are we can better decide when - or even if - we use each technique. Once we understand the relationship of technique to suitable circumstance, we can train in technique selection. Once we understand the specific applications of technique, we can train in heuristic-like shortcuts (for example, on military battlefield scenarios we use an ACH-like technique for selecting likely and dangerous COAs from assessed enemy capability).

Without getting better – without understanding what works and what does not, and when, and without discarding inappropriate tools – we are unlikely to produce the assessments that our decision makers need, and are likely to fall far behind opponents who are more rigorous in examining their working practice.

Q1s: What has been your experience with applying the analytic techniques you were taught, in practice? Has there ever been a moment when one 'saved the day'? Which ones do you find useful, and which ones not, and why?

Q2s: what are the factors that you think are important when testing? We have: timeliness, effort to apply, effort to learn, extra costs (equipment etc), rigour, accuracy, audit, objectivity, reliability.

--

*Martin Hill is an Army Reserve Corporal, teaching analytic techniques at Chicksands, with past roles in light cavalry and combat engineering, and an operational tour as infantry. He is a Visiting Researcher at Cranfield University, the Defence Academy UK, studying Knowledge Distribution. He is a freelance software engineer who updates large, complex, legacy software systems for large organisations.*

[1] M. Hill and J. Salt, “Life and Death Decisions using Sparse Unreliable Evidence,” in *ECIME*, 2010, vol. 53, no. 9.

[2] Development Concepts and Doctrine Centre - MOD, *Understanding and Intelligence Support to Joint Operations (JDP 2-00 JWP 2-00)*. UK MoD, 2011.

[3] S. Kent, *Words of Estimative Probability*. CIA 2007, 1964.

[4] R. H. Pherson and R. Heuer, *Structured Analytic Techniques for Intelligence Analysis*. CQ Press, 2010.

[5] DIA, *Quick Wins for Busy Analysts*. 2013.

[6] CFIC, *Aide Memoire on Intelligence Analysis Tradecraft*, no. August. 2015.

[7] M. Whitesmith, “Efficacy of ACH in mitigating bias,” *Intell. Natl. Secur.*, 2019.

[8] K. Possin, “CAT scan: A critical review of the critical-thinking assessment test,” *Informal Log.*, vol. 40, no. 3, pp. 489–508, 2020.